Balanced parentheses in NL texts : a useful cue in the syntax/semantics interface

Gabriel G. Bès, Veronica Dahl

Balanced parentheses in NL texts: a useful cue in the syntax/semantics interface*

Gabriel G. Bès†
Université Blaise-Pascal
GRIL

Veronica Dahl‡
Simon Fraser University
Computing Sciences Department

Abstract

Balanced parentheses on text sentences can be obtained from information on particular morphemes – the introducers – and on inflected verbal forms. From balanced parentheses, a partial graph of the sentence in the semantics interface can be deduced, along with other information. The hypothesis and its expression with CHR constraints are presented.

1 The basic hypothesis

Many formal languages use parentheses, left ones (*lp*) and right ones (*rp*). They are balanced: at the end of a well formed expression $N(lp) = N(rp)$ (where *N: number*), and, at any point of it, $N(lp) \leq N(rp)$.

Parentheses are well identified objects in formal languages. Classified under the label of "auxiliary symbols" they do not have intrinsic semantic value, but they are crucially important for the specification of operators domains. In Montague Grammar (Montague, 1974), their expressive power is even greater: indexed parentheses encode the syntactic operation from which they follow.

Parentheses are widely used in formal or quasi formal syntactic representations. But, as pointed out by Hintikka (Hintikka, 1994), they are not "natural" objects. They belong to the syntactic machinery of the metalanguage used to describe NL expressions, and as such, their use can freely change from one machinery to another, even if they remain balanced.

But not all formal languages need parentheses as auxiliary symbols. The polish notation of first order logic does not require them. Our central hypothesis is a kind of an answer to Hintikka's challenge objection: balanced parentheses can indeed be deduced – and not stipulated – from an adequate analysis of NL expressions. Furthermore, they lead to a partial graph, which can be used as an important cue in the syntax/semantic interface.

Balanced parentheses can be obtained from an adequate analysis of a subset of grammatical morphemes such as French *si, que,...*, the introducers, and inflected verbal forms, inflected chunks (e.g. *a lu, lui a donné*, or inflected verbs (e.g. *aimait, parle*). Metaphorically, they allow to jump to the roof of a sentence from poor information on local marks in its foundations.

2 From local information to the partial graph

The balanced parentheses hypothesis can be illustrated by the following (i), analyzed by the subsequent (ii) to (v).

i Si les parents s'étaient mis d'accord hier et avaient bien connu la réglementation, les bureaucrates à qui ils se sont adressés aujourd'hui ne leur auraient pas répondu que c'était impossible, ils auraient dû présenter leur dossier autrement.

With respect to (i), it is possible to say that there are inflected nuclear verbal phrases (*vnfl*), as *se sont adressés*, that in one case, two *vnfl*s coordinate (*s'étaient mis d'accord* and *avaient bien connu*), that this verbal coordination is the verbal form of the conditional sentence, that the verb form of the root sentence is *ne leur*

*Authors in alphabetical order. Thanks are given to Caroline Hagège for extended and enlightening discussions in the preliminairies of this work, and to François Trouilleux for his commments.

†Gabriel.Bes@univ-bpclermont.fr; 34 Ave. Carnot, F 63037 Clermont-Fd cedex.

‡veronica@sfu.ca; 8888 University Dr. Burnaby B.C. V5A 1S6 Canada.

auraient pas répondu, that the whole sentence with *auraient dû* as verbal form is coordinated to the root sentence.

Futhermore, it is possible to say that there are morphological expressions, simple (as *si*) or complex (as *à qui*), which flag a coming *vnfl*; these are the *introducers*. For instance *se sont adressés* is introduced by *à qui*, *auraient dû* by the nominative form *ils*. The first *vnfl* of the coordinated verbal form of the conditional sentence is introduced by *si* and the *vnfl* of the root sentence is introduced by a hidden *il (initial limit)*, assumed as first element of any expression, as *fp (final point)* is the final one.

If we introduce an *lp* at the left of *il* and an *rp* at the right of *fp*, associate an *lp* to each introducer and an *rp* to each introduced *vnfl* not coordinated with another *vnfl*, and if we associate an *lp* to the first *vnfl* of coordinated *vnfl*s and two *rp* to the right of the last coordinated *vnfl*, balanced parentheses on the whole sentence are obtained.

Thus, from (i), the following (ii) is obtained. In (ii), '-' joins single expressions in (i), obtaining chunk expressions which are computed, with respect to position and tags, as the simple ones. A position is assigned in (ii) to each expression jointly with a tag from a very restricted vocabulary V = *{int, v, v1, v2, il, fp, ot}*.

Besides *il* and *fp*, presented earlier, *int* in V is associated to introducers, *v* is associated to *vnfl*s not immediately preceeded either by ',' or by a coordination form, *v1* is associated to *vnfl*s immediately preceeded by ',' *v2* is associated to *vnfl*s immediately preceeded by a coordination form (as *et, ou...*) not preceeded by a ',', and *ot (other)* is associated to any expression which is not associated to one of the previous tags. *INT* will spell both *int* and *il*.

ii ((il$_{<0,il>}$ (Si$_{<1,int>}$ les$_{<2,ot>}$ parents$_{<3,ot>}$ (s'-étaient-mis-d'accord$_{<4,v>}$ hier$_{<5,ot>}$ et-avaient-bien-connu$_{<6,v2>}$)) la$_{<7,ot>}$ réglementation$_{<8,ot>}$,$_{<9,ot>}$ les$_{<10,ot>}$ bureaucrates$_{<11,ot>}$ (à-qui$_{<12,int>}$ ils$_{<13,ot>}$ se-sont-adressés$_{<14,v>}$) aujourd'hui$_{<15,ot>}$ ne-leur-auraient-pas-répondu$_{<16,v>}$) (que$_{<17,int>}$ la$_{<18,ot>}$ chose$_{<19,ot>}$ était$_{<20,v>}$) impossible$_{<21,ot>}$,$_{<22,ot>}$ (ils$_{<23,int>}$ auraient-dû$_{<24,v>}$) présenter$_{<25,ot>}$ leur$_{<26,ot>}$ dossier$_{<27,ot>}$ autrement$_{<28,ot>}$ pf)

If we eliminate NL expressions, leaving only parentheses, tags from V and positions, we obtain the more perspicuous (iii) or (iv). In (iv), '...', spelling *intervals*, substitutes for *ot*s.

iii ((int$_0$ (int$_1$ ot$_2$ ot$_3$ (v$_4$ ot$_5$ v2$_6$)) ot$_7$ ot$_8$ ot$_9$ ot$_{10}$ ot$_{11}$ (int$_{12}$ ot$_{13}$ v$_{14}$) ot$_{15}$ v$_{16}$) (int$_{17}$ ot$_{18}$ ot$_{19}$ v$_{20}$) ot$_{21}$ ot$_{22}$ (int$_{23}$ v$_{24}$)ot$_{25}$ ot$_{26}$ ot$_{27}$ ot$_{28}$ fp)

iv ((int$_0$ (int$_1$...(v$_4$...v2$_6$))...(int$_{12}$...v$_{14}$) ...v$_{16}$)(int$_{17}$...v$_{20}$) ...(int$_{23}$ v$_{24}$)...fp)

In (iv), besides intervals, there are *INT*s and *vnfl*s (i.e. *v, v2*) each in some position *0* to *n*. These relations can be expressed by pairs *<i, j>*, where $i \neq j$ and $i, j \geq 0$. These pairs will be assigned to different sets.

By general convention, *q* is the position of the *vnfl* introduced by some *INT*, i.e. the non coordinated *vnfl* associated to the ')' which closes the associated '(', or the first *vnfl* in a coordinated chain of *vnfl*s, coordinated chain which closes the *INT*. If *INT* = *il*, we express the relation by *<q, 0>*, if *INT* ≠ *il* and in position *p*, by *<p, q>*.

Closing pairs (both *<q, 0>* and *<p, q>*) are in the set *Cl[osing]*. From *<q, 0>* ∈ *Cl* we can deduce that *q* ∈ *R[oot]*, where *R* is either an empty set (see §3.3) or a singleton set with the position of the root *vnfl* as member.

Coordinated *vnfl*s in verbal phrases, which are in chains *vnfl*$_{w1}$ ... *vnfl*$_{wn}$, are denoted by coordination pairs *<w1, wi>*, where *wi* ≠ *w1*. Coordination pairs of verbal phrases are in the set *C-sv*. A *vnfl* in position *q* and closing some *int* ≠ *il*, can be the verbal form of a sentence coordinated to the root sentence (e.g. *auraient-dû*$_{<24,v>}$ in (ii)). In this case, we write *<q, 0>* and the pair belongs to the *C-r* set. With these conventions, from (iv) we obtain (v).

(v) Cl = {<16, 0>, <1, 4>, <12, 14>, <17, 20>, <23, 24>}
C-sv = {<6, 4>}
R = {16}
C-r = {<24, 0>}

Cl, *C-sv* and *C-r* being sets of pairs, from the union of them it is possible to deduce a partial graph, positions in the input being its vertices.

The parsing system that obtains the elements in (v), given the input expression in (i), is organized in Modules I and II. Module I, succinctly presented here, has as input a chain of Ascii codes of NL texts, associated to one or more

sentences, and obtains representations as in (ii), with one or more segmented and enumerated sentences. An interface obtains (iii) from (ii). The challenge of Module I is the disambiguation of expressions such as *si* or *la juge* which can be or not *int*s or *vnfl*s, respectively. It is obtained by exploring local contexts. Module II is expressed in two different ways. There is an algorithm (*Algof-c*) which from (iii) obtains (v). The other way is a plain declarative one, making use of CHR constraints.

3 CHR constraints

CHR (Frühwirth and Abdennadher, 2003) is a very powerful multiset rewriting language. Constraints, viewed as pieces of partial information, are formalized as distinguished, predefined predicates in first-order predicate logic.

A constraint program successively generates constraints as it runs, until a solution is found to the problem or no more constraints can be generated. Rules describe how to generate new constraints from those already generated. For instance, we can view symbols in a grammar as constraints upon word boundaries in an input string. Thus *an interesting morning* could be parsed by CHR rules such as:

```
(1) an(X,Y) ==> det(X,Y).
(2) interesting(X,Y) ==> adj(X,Y).
(3) morning(X,Y) ==> noun(X,Y).
(4) start ==> an(1,2),
interesting(2,3), morning(3,4).
(5) det(X,Y), adj(Y,Z), noun(Z,W)
==> np(X,W).
```

The contiguity and order of the *det, adj* and *noun* are ensured by the word boundaries; e.g., the *det* ends where the *adj* starts, at point *Y*. *GC* is the set of all the generated constraints derivable from the input string defined through (1) to (4). It will be generated upon the query:`?- start`. In *GC*, we can then select those that solve the problem we are interested in (in this case `np(1,4)`, which tell us our string analyses into a `np`).

3.1 CHR constraints and grammar rules

CHR rules can directly mirror grammar rules: (5) in §3 mirrors `np` → `det adj noun`, assuming contiguity between `det adj noun`. CHR rules generate constraints, bottom up and left to right, implementing grammar rules.

Given a structure $\ldots vnfl_i \ldots vnfl_j \ldots$, not all *vnfl* (i.e. *v, v1, v2*) can instantiate $vnfl_i$ or $vnfl_j$. For instance, assuming *h* to the left of *i*:

(1) If $vnfl_i$ = *v2* and it coordinates with $vnfl_h$ = *v or v1*, then $vnfl_j \neq$ *v2*.
(2) If $vnfl_i$ = *v2* and it does not coordinate with $vnfl_h$, or if it coordinates to a $vnfl_h$ = *v2*, then $vnfl_j$ may be a *v2* .

The grammar which mirrors CHR Rules is thus a grammar of type 1 in the Chomsky hierarchy[1].

The format of the input of CHR rules is

il $\ldots$ x_1 $\ldots$ x_n $\ldots$ x_m $\ldots$ fp

where x_i is either an *int* or a *vnfl*, and '$\ldots$' is, here, either an interval or *e(mpty)*. The basic challenge of CHR rules is to specify the constraints in *GC* from which arcs in the partial graph can be obtained.

3.2 Arcs from constraints

The whole set *GC* is not needed for obtaining arcs. *GC* is thus the domain of partial functions specifiying pairs of the resulting graph in its range. As an illustration on verbal coordination, consider *a regardé, regarde et regardera ce tableau*, with (vi) as its CHR-rules input.

(vi) `v(1,2), v1(2,3), v2(3,4), ot(4,5), ot(5,6).`

Several grammar rules specify different types of verbal coordination, two of them underlying the specification of the C-sv set related to (vi):

vOc1 → *v v1*, which coordinates *v1* to *v* obtaining *vOc1*

vOc → *v v1 v2*, which coordinates *v1* and *v2* to *v* obtaining *vOc*

The CHR rules obtain the *GC* (vii) from (vi).

(vii)
```
v(1,2), v0(1,2), v1(2,3),
v1R(2,3), v0c1(1,3), v1NT(2,3),
cv(1,2), v2(3,4), v2R(3,4),
v1c(2,4), v0c(1,4), ot(4,5),
ot(5,6), otR(4,6), cv(2,5),
cv(1,5), ?  yes
```

All the informations in (vii) are not needed for obtaining elements in *C-sv*. For instance, intervals, expressed by *otR*, are not significant for

[1] The grammar and the CHR-rules program can be provided on demand.

the extraction of the partial graph. Among the partial functions with *GC* as domain, we have *F1* and *F2*. Given (vii), <2,1>, <3,1> ∈ C-sv are obtained by *F1* and *F2*, respectively.

F1 : v0c1(X,Z), v(X,Y), v1R(Y,Z) ∈ GC
→ <(Z-1),X> ∈ C-sv

F2 : v0c(X,W), v0c1(X,Z), v2(Z,W) ∈ GC
→ <Z,X> ∈ C-sv

Another example illustrates the obtention of the *Cl* set. Consider the embedded sentences *(dit)* que_1 la_2 $fille_3$ que_4 $Jacques_5$ $a\text{-}regardée_6$ $est\text{-}partie_7$ with (viii) as its input to CHR rules.

(viii) `int(1,2),..., int(4,5),..., v(6,7), ...,v(7,8)`

(1) and (2) in (ix) compact several grammar rules. In (1), *X* is *cf, (constituant fermé)*, or *e*, and *iNT* rewrites as *int ot**, *ot** expressing intervals or *e*. In (2), *Y* explicit contextual restrictions (see §3.1), while *Z* rewrites *cv (complexe verbal)*, obtaining terminal strings (a *vnfl* or a *vnfl* coordination) with an initial *vnfl*.

(ix) (1) $cf \rightarrow iNT\ X\ cv$
(2) $cv/Y \rightarrow Z\ ot*$

There are CHR rules which mirror (ix). In (x), (1) is a subset of the GC obtained by them, (2) is the partial function *F3*. Given (viii), (1) and (2), (3) is obtained.

(x) (1) {cf(x,y), cv(y,z)}
(2) F3 : cf(X,Y), cv(Y,Z) ∈ GC
→ <X,Y> ∈ Cl
(3) <4, 6>, <1, 7> ∈ Cl

3.3 Deduced information on intervals

From arcs obtained by partial functions from *GC*, besides the possibility of expressing the semantic representation of verbal coordination, other interesting informations can be deduced. For instance, if $<0, p> \notin Cl$, then $R = \{\}$, and it is likely that the expression will be a nominal phrase, even with one or more embedded relatives. Furthermore, deduced parentheses associated to particular input symbols specify intervals with inherent restrictions, as in (ix).

(xi) (1) (int_i ... (int_j
(2) (int_i ... VFORM, where VFORM is vnfl_j) or ((vnfl_j
(3) vnfl_i ... vnfl_j

A nominal phrase in (1) can be the subject of a *vnfl* in some position to the right of position *j*, while that is not the case with (2) or (3).

4 Ongoing work and discussion

Ongoing work relates to the extension of verbal coordinations, to the improvement of the expressive power of the grammar (today with less expressive power than *Algof-c*, see §3) and its CHR implementation and, last but not the least, to the evaluation in effective texts of the underlying linguistic hypothesis, knowing beforehand that neither all *int*s or all *vnfl*s can be obtained by Module I, nor all verbal coordination be handled by Module II.

Even if we know this, we claim that, in general, from poor and local information obtained by Module I, thanks to the deductible character of balanced parentheses in NL texts, it is possible to obtain a partial graph of the whole sentence, with good and effective approximations in the NL-software engineering domain. From this, besides verbal coordination, it is possible to deduce, in turn, restrictions on intervals, which will reduce the parsing research space[2].

References

Luísa Coheur, Nuno Mamede, and Gabriel G. Bès. 2003. Asdecopas: a syntactic-semantic interface. In *EPIA'03 Workshop on Natural Language and Text Retrieval*, Evora (Portugal).

T. Frühwirth and S. Abdennadher. 2003. *Essentials of Constraint Programming.* Springer Verlag.

Jaako Hintikka and Gabriel Sandu. 1997. Game theoretical semantics. In van Johan Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 361–410. Elsevier.

Jaako Hintikka. 1994. *Fondements d'une théorie du langage.* PUF.

Richard Montague. 1974. Universal grammar. In Richard Thomason, editor, *Formal Philosophy, selected papers of Richard Montague*, pages 222–246. Yale University Press.

[2]The conjecture is that the analysis of intervals will obtain new arcs and vertices, and from them an oriented graph on the sentence, from which, in turn, semantic functions will obtain the semantic representation, cf. (Coheur et al., 2003), but, following (Hintikka, 1994) and (Hintikka and Sandu, 1997), we do not claim that from balanced parentheses scopes of quantifiers can be deduced.